

Multicovariate-adjusted regression models

D. V. NGUYEN*[†] and D. ŞENTÜRK[‡]

[†]University of California Davis, USA

[‡]Pennsylvania State University, USA

(Received 26 July 2006; final version received 27 April 2007)

We introduce multicovariate-adjusted regression (MCAR), an adjustment method for regression analysis, where both the response (Y) and predictors (X_1, \dots, X_p) are not directly observed. The available data have been contaminated by unknown functions of a set of observable distorting covariates, Z_1, \dots, Z_s , in a multiplicative fashion. The proposed method substantially extends the current contaminated regression modelling capability, by allowing for multiple distorting covariate effects. MCAR is a flexible generalisation of the recently proposed covariate-adjusted regression method, an effective adjustment method in the presence of a single covariate, Z . For MCAR estimation, we establish a connection between the MCAR models and adaptive varying coefficient models. This connection leads to an adaptation of a hybrid backfitting estimation algorithm. Extensive simulations are used to study the performance and limitations of the proposed iterative estimation algorithm. In particular, the bias and mean square error of the proposed MCAR estimators are examined, relative to a baseline and a consistent benchmark estimator. The method is also illustrated with a Pima Indian diabetes data set, where the response and predictors are potentially contaminated by body mass index and triceps skin fold thickness. Both distorting covariates measure aspects of obesity, an important risk factor in type 2 diabetes.

Keywords: covariate adjusted regression; local polynomial regression; multiplicative effect; varying-coefficient models

2000 Mathematics Subject Classifications: 62G08; 62J02; 62J05

1. Introduction and example

Adjusting for anthropometric measurements, such as body mass index (BMI) and/or other measures of body configuration, is common in medical or health related studies because they are distorting variables that affect the primary variables of interest. For example, in a study involving haemodialysis patients, a primary outcome variable is an elevated plasma fibrinogen level [1, 2]. Fibrinogen is a protein found in blood plasma and it is a risk factor for cardiovascular disease in haemodialysis patients. It is of interest to examine the relationship between fibrinogen concentration and other predictors, such as serum transferrin protein level. However, both primary variables of interest, fibrinogen and transferrin protein levels, are known to depend on body mass index (defined as $\text{weight}/\text{height}^2$), which exerts a distorting

*Corresponding author. Email: ucdnguyen@ucdavis.edu

effect on the protein measurements. A common approach to adjust for the distorting covariates, like BMI, is to normalise the primary variables of interest by simply dividing (by BMI). Note that this adjustment by division implies that the assumed contamination is of a multiplicative form. To set notations, let \tilde{Y} , \tilde{X} and Z denote the observed fibrinogen concentration, serum transferrin level and covariate BMI, respectively. Using these notations, the adjusted primary variables that are thought to be free from the distorting effect of BMI are,

$$Y = \frac{\tilde{Y}}{Z} \quad \text{and} \quad X = \frac{\tilde{X}}{Z}. \quad (1)$$

The basic motivation for the above adjustment is to obtain normalised versions of the observed primary variables by removing the distorting covariate effects, so that the measurements are comparable across patients. One, then, targets the regression relationship between Y and X , free from the effects of Z .

Motivated by the practice of multiplicative adjustments for covariate effects, Şentürk and Müller [2] proposed a more flexible multiplicative adjustment procedure for regression models. They directly modelled the distortion through *unknown functions* of Z . More precisely, their adjustment method models the underlying response and predictor variables of interest as

$$Y = \frac{\tilde{Y}}{\psi(Z)}, \quad X_1 = \frac{\tilde{X}_1}{\phi_1(Z)}, \dots, X_p = \frac{\tilde{X}_p}{\phi_p(Z)}, \quad (2)$$

where $\psi(\cdot)$, $\phi_1(\cdot)$, \dots , $\phi_p(\cdot)$ are unknown smooth contaminating functions of a single covariate, Z . Allowing for the unknown contaminating functions in equation (2) is an appealing aspect, from a practical point of view. This is because, in practice, the precise nature of the multiplicative relationships between the distorting covariate and the primary variables of interest is unknown. Lacking this precise knowledge, the naive practice of dividing by the covariate in equation (1) or equivalently assuming $\psi(Z) \equiv Z$ and $\phi_r(Z) \equiv Z$ in equation (2) imposes unnecessarily rigid constraints on the form of the data contamination. Assuming a more general contamination (2), Şentürk and Müller (2005) target the parameters from the underlying regression model, $E(Y) = \gamma_0 + \gamma_1 X_1 + \dots + \gamma_p X_p$, based on the contaminated observations (\tilde{Y} and $\{\tilde{X}_r\}_{r=1}^p$).

In many applications, there are multiple distorting covariates that simultaneously affect the primary variables of interest. Thus, in this paper, we explore an important generalization of the above method (for a single covariate Z) to the case of multiple distorting covariates, namely $\mathbf{Z}^T = (Z_1, \dots, Z_s)$. This generalisation allows for additional modelling flexibility by accommodating multiple covariates. We model the distortion through unknown functions of an unspecified linear combination of the covariates, $\boldsymbol{\eta}^T \mathbf{Z}$, where $\boldsymbol{\eta}^T = (\eta_1, \dots, \eta_s)$ are unknown coefficients to be estimated from the data. Thus, our adjustment method models the underlying response and predictor variables of interest as

$$Y = \frac{\tilde{Y}}{\psi(\boldsymbol{\eta}^T \mathbf{Z})}, \quad X_1 = \frac{\tilde{X}_1}{\phi_1(\boldsymbol{\eta}^T \mathbf{Z})}, \dots, X_p = \frac{\tilde{X}_p}{\phi_p(\boldsymbol{\eta}^T \mathbf{Z})}. \quad (3)$$

Under this general contamination by multiple covariates, \mathbf{Z} , we target the underlying parameters in the linear regression model of Y on $\{X_r\}_{r=1}^p$, which is not directly observable. A data illustration, provided in section 5, examines the underlying relationship between plasma glucose concentration and (diastolic) blood pressure, adjusted for an obesity index based on BMI and triceps skin fold thickness. Both distorting covariates are related to obesity, which is an important modifiable risk factor of complications resulting from type 2 diabetes (Diabetes Mellitus Type II). We note that under models (1) and (3) the interpretation of Y and X_r is

the same; that is, they are the parts of \tilde{Y} and \tilde{X}_r that are free of the effects of U (where $U = Z$ in the univariate case (1) and $U = \beta^T \mathbf{Z}$ in the multivariate case (3)). In the above data example, these are the obesity index adjusted plasma glucose concentration and diastolic blood pressure.

Adjusting for multiple distorting covariates poses many additional challenges over the case of a single known covariate (2). For example, the computational complexity increases substantially because both the unknown contaminating functions ($\psi(\cdot)$ and $\phi_r(\cdot)$, $r = 1, \dots, p$) and their index $\boldsymbol{\eta}^T \mathbf{Z}$ need to be estimated simultaneously in order to obtain the underlying regression parameters ($\{\gamma_r\}_{r=0}^p$). This is feasible through the use of the one-dimensional index or linear combination, $\boldsymbol{\eta}^T \mathbf{Z}$, which helps deal with the curse of dimensionality.

Also, the proposed adjustment method adjusts for multiplicative distortion, as well as additive (*i.e.*, $\tilde{Y} = Y + \psi(\boldsymbol{\eta}^T \mathbf{Z})$, $\tilde{X}_r = X_r + \phi_r(\boldsymbol{\eta}^T \mathbf{Z})$) and no distortion (*i.e.*, $\psi(\boldsymbol{\eta}^T \mathbf{Z}) = \phi_r(\boldsymbol{\eta}^T \mathbf{Z}) = 0$ under additive and $\psi(\boldsymbol{\eta}^T \mathbf{Z}) = \phi_r(\boldsymbol{\eta}^T \mathbf{Z}) = 1$ under multiplicative distortions as in [2]). This is mainly due to the identifiability conditions given in detail in section 2. If additive distortion is known or assumed, then it can be handled with partial regression methods. However, there is no adjustment method available for handling multiplicative and multivariate distortion. The proposed multivariate-adjusted regression (MCAR) adjustment handle these types of distortion automatically without prior specification of the exact type (*i.e.*, multiplicative, additive or no distortion).

The paper is organised as follows. We describe the formulation of MCAR and establish an important relationship between the MCAR model and the adaptive varying coefficient model of the form $E(\tilde{Y}|\tilde{\mathbf{X}}, \mathbf{Z}) = \sum_{r=0}^p \beta_r(\boldsymbol{\eta}^T \mathbf{Z}) \tilde{X}_r$ ($\tilde{X}_0 \equiv 1$) in section 2. This result leads us to adapt a hybrid backfitting algorithm (section 3) to estimate the unknown varying functions $\beta_r(\cdot)$ and the coefficient vector $\boldsymbol{\eta}$ simultaneously, which are needed for the estimation of γ . Simulation studies in section 4 examine the performance and limitations of the proposed method. We (a) compare the bias and mean square error of the proposed iterative MCAR estimator to a baseline and a consistent benchmark estimator, (b) assess the sensitivity of the algorithm to the starting values for $\boldsymbol{\eta}$ and (c) examine the sample size effect. The MCAR method is illustrated in section 5 with a data set on diabetes in females with Pima Indian heritage. We conclude with a discussion in section 6.

2. Multivariate-adjusted regression models

Consider the problem of estimating the parameters $\{\gamma_r\}_{r=0}^p$ of the model

$$Y_i = \gamma_0 + \sum_{r=1}^p \gamma_r X_{ir} + e_i, \quad (4)$$

where Y_i and $\{X_{ir}\}_{r=0}^p$ are the response and predictor values corresponding to the i th subject, respectively. The error variable e_i has $E(e_i) = 0$ and $\text{var}(e_i) = \sigma^2$. We only observe n copies of distorted predictor and response data, $\{\tilde{Y}_i, \tilde{\mathbf{X}}_i\}_{i=1}^n$, along with multiple covariates \mathbf{Z}_i , where

$$\tilde{Y}_i = \psi(\boldsymbol{\eta}^T \mathbf{Z}_i) Y_i \quad \text{and} \quad \tilde{X}_{ir} = \phi_r(\boldsymbol{\eta}^T \mathbf{Z}_i) X_{ir}, \quad r = 1, \dots, p. \quad (5)$$

Some constraints on the unknown smooth distortion functions are still needed for the identifiability of the estimation problem. A set of reasonable constraints for $\psi(\cdot)$ and $\{\phi_r(\cdot)\}$ is implied by the natural assumption that the mean distorting effect should correspond to no

distortion [2], i.e.,

$$E\{\psi(\eta^T \mathbf{Z})\} = 1 \quad \text{and} \quad E\{\phi_r(\eta^T \mathbf{Z})\} = 1. \quad (6)$$

It is assumed that $\{(\mathbf{X}_i, \mathbf{Z}_i^T, e_i)\}_{i=1}^n$ are independent and identically distributed, where \mathbf{X} , e and \mathbf{Z} are mutually independent. We refer to the multiplicative distortion model, described by equations (4)–(6), as the MCAR model.

Our main objective is estimation of the underlying regression parameters γ , given the contaminated observations and distorting covariates. Towards this objective, we first establish an important relation between the estimation problem under MCAR and estimation under the adaptive varying coefficient model. Given the contaminated observations and the distorting covariates, a regression of \tilde{Y} on $\{\tilde{X}_r\}_{r=1}^p$ leads to the following regression relation,

$$E(\tilde{Y}|\tilde{\mathbf{X}}^T, \mathbf{Z}) = \psi(\eta^T \mathbf{Z}) E \left\{ \gamma_0 + \sum_{r=1}^p \gamma_r X_r + e|\phi_1(\eta^T \mathbf{Z})X_1, \dots, \phi_p(\eta^T \mathbf{Z})X_p, \mathbf{Z} \right\}.$$

Further simplifications, using equation (5) and mutual independence of $(e, \mathbf{Z}$ and $X_r)$, give

$$E(\tilde{Y}|\tilde{\mathbf{X}}^T, \mathbf{Z}) = \beta_0(\eta^T \mathbf{Z}) + \sum_{r=1}^p \beta_r(\eta^T \mathbf{Z}) \tilde{X}_r, \quad r = 1, \dots, p, \quad (7)$$

where

$$\beta_0(\eta^T \mathbf{Z}) = \gamma_0 \psi(\eta^T \mathbf{Z}) \quad \text{and} \quad \beta_r(\eta^T \mathbf{Z}) = \gamma_r \{\psi(\eta^T \mathbf{Z})/\phi_r(\eta^T \mathbf{Z})\}. \quad (8)$$

Thus, the MCAR model leads to the following adaptive varying coefficient model [3,4], $\tilde{Y} = \beta_0(\eta^T \mathbf{Z}) + \sum_{r=1}^p \beta_r(\eta^T \mathbf{Z}) \tilde{X}_r + \varepsilon$, with $\varepsilon \equiv \psi(\eta^T \mathbf{Z})e$. The varying coefficient functions, $\beta_r(\cdot)$ ($r = 1, \dots, p$), are proportional to the quotient of the original distorting functions, $\{\psi(\cdot)/\phi_r(\cdot)\}$; and the intercept function, $\beta_0(\cdot)$, is proportional to $\psi(\cdot)$. The constants of proportionality are the underlying regression parameters, $\{\gamma_r\}$, of interest.

Varying coefficient models [5,6] are popular in many application areas. The literature includes, among others, [7] on functional data analysis and [8] and [9] on nonlinear time series. Some approaches to estimation in varying coefficient models for independent and identically distributed data are described in [10–12]. The literature related to the adaptive varying coefficient model (7), where the index is unknown, includes refs. [3,4,13,14]. Note also that the adaptive varying coefficient model (7) differs from the primary model considered in [4], in that the index variable U is not a linear combination of the predictors $\tilde{\mathbf{X}}$. The index for MCAR is a linear combination of the distorting covariates, namely $U = \eta^T \mathbf{Z}$, where the covariates \mathbf{Z} are different from the predictors $\tilde{\mathbf{X}}$. As will be detailed in the next section, we adopt a similar approach as in [4] to first estimate the unknown functions $\{\beta_r(\cdot)\}$ and the direction, η , simultaneously. We then target the regression parameters, γ , by weighted averages of the $\{\beta_r(\cdot)\}$ estimates.

Before we proceed to the estimation, we note some important distinctions between the proposed MCAR models and varying coefficient models. MCAR can be viewed as a latent variable model, where the underlying linear model has general multiplicative distortion structures. Furthermore, the connection between MCAR and the varying coefficient model (7) is a convenient estimation tool that is used to recover the underlying relationship between the latent response and the predictors, free from the distorting effects of $\{Z_1, \dots, Z_s\}$. Therefore, one of the main distinctions between MCAR and varying coefficient models is that the distorting variables, namely the Z 's (or the single index $U = \eta^T \mathbf{Z}$), are considered ‘nuisance’ variables under MCAR. On the contrary, these are of main interest in a varying coefficient model analysis.

3. Estimation of the underlying regression parameters

To motivate our estimation algorithm for the underlying regression parameters, $\{\gamma_r\}$, in the MCAR model, let us first consider the case where η is known. With η known, the form of the distorting covariates $U \equiv \eta_1 Z_1 + \cdots + \eta_s Z_s$ is completely observed. This is equivalent to having a single observable covariate U . Thus, the adaptive varying coefficient model, given by equation (7) reduces to a standard varying coefficient model, $\tilde{Y} = \beta_0(U) + \sum_{r=1}^p \beta_r(U) \tilde{X}_r + \epsilon$. We emphasise that the uncertainty due to η is now completely eliminated. Under this situation, the following weighted-average estimators are consistent for the underlying regression parameters $\{\gamma_r\}$,

$$\hat{\gamma}_{0*} = n^{-1} \sum_{i=1}^n \hat{\beta}_0(U_i) \quad \text{and} \quad \hat{\gamma}_{r*} = \frac{1}{\bar{X}} \sum_{i=1}^n \frac{1}{n} \hat{\beta}_r(U_i) \tilde{X}_{ir}, \quad (9)$$

where $\bar{X}_r = n^{-1} \sum_{i=1}^n \tilde{X}_{ir}$, $U_i = \eta^T \mathbf{Z}_i$ and $\{\hat{\beta}_r(\cdot)\}$ are local linear estimators of the varying coefficient functions. Explicit formulas for $\{\hat{\beta}_r(\cdot)\}$ are given in the next section. The consistency of $\hat{\gamma}_{r*}$, $r = 0, \dots, p$ follow from the consistency result for the case of a univariate covariate given in [15]. The estimators in equation (9) provide a benchmark for systematically studying the performance of the proposed iterative MCAR estimator, where η will not be known. We also note that the special weighting scheme utilised in equation (9) was originally proposed in [2] and it was designed to eliminate the impact of the distorting functions. However, they used a binning approach rather than local linear estimators for $\{\beta_r(\cdot)\}$, which was recently proposed in [15]. The later approach equation (9) will be used here, although both approaches are equivalent for large sample sizes.

When η is unknown, we also need to estimate $\eta^T \mathbf{Z}_i$ for the i th subject ($i = 1, \dots, n$). Given the consistency of equation (9) and the established relationship in section 2, we adopt a hybrid backfitting algorithm similar to the one proposed for adaptive varying coefficient models [4] to simultaneously estimate $\eta^T \mathbf{Z}$ and $\beta_r(\eta^T \mathbf{Z})$. Briefly, the algorithm for estimating the varying coefficient functions consists of two main steps. Step (1) Given an initial vector (or starting value) η_0 , estimate the varying functions $\beta_r(\cdot)$, using local linear regression. These are the initial local linear estimates for $\{\beta_r(\cdot)\}$. Step (2) Next, fixing $\{\beta_r(\cdot)\}$ (*i.e.*, obtained based on η_0), one can search for or update η using a one-step Newton–Raphson scheme. Steps (1) and (2) are repeated/iterated until convergence occurs based on a mean squared error criterion. Let $\hat{\eta}$ be the value at convergence and $\{\hat{\beta}_r(\hat{\eta}^T \mathbf{Z})\}_{r=0}^p$ be the corresponding local linear estimates of the coefficient functions. Using these estimates in combination with the distortion eliminating weights, we arrive at the MCAR estimators of the underlying regression parameters, analogous to the consistent benchmark estimators (9),

$$\hat{\gamma}_0 = n^{-1} \sum_{i=1}^n \hat{\beta}_0(\hat{\eta}^T \mathbf{Z}_i) \quad \text{and} \quad \hat{\gamma}_r = \frac{1}{\bar{X}} \sum_{i=1}^n \frac{1}{n} \hat{\beta}_r(\hat{\eta}^T \mathbf{Z}_i) \tilde{X}_{ir}, \quad (10)$$

3.1 Initial local linear regression estimators for $\beta_r(\cdot)$

For a given η , the varying coefficient function $\beta_r(\cdot)$ can be approximated based on local linear modelling as $\beta_r(\eta^T \mathbf{Z}) \approx \beta_r(u) + \beta'_r(u)(\eta^T \mathbf{Z} - u)$, $r = 0, 1, \dots, p$, for $\eta^T \mathbf{Z}$ in a neighborhood of u [16]. The $\beta'_r(\cdot)$ denotes the derivative of $\beta_r(\cdot)$. The local linear estimators of $\{\beta_r(\cdot)\}$

are obtained by minimising the sum

$$\sum_{i=1}^n \left[\tilde{Y}_i - \sum_{r=0}^p \left\{ b_r + c_r (\boldsymbol{\eta}^T) \tilde{X}_{ir} \right\} \right]^2 K_h(\boldsymbol{\eta}^T \mathbf{Z}_i - u),$$

with respect to $\{b_r, c_r\}$ and for a specified kernel function K with bandwidth h . This minimisation is a weighted least squares problem, so the local linear estimators follow directly from least squares theory. Let $\hat{\beta}_r(u) = \hat{b}_r$, $\hat{\beta}'_r(u) = \hat{c}_r$ and $\hat{\boldsymbol{\alpha}} \equiv (\hat{b}_0, \dots, \hat{b}_p, \hat{c}_0, \dots, \hat{c}_p)^T$. The local linear regression estimates $\hat{\boldsymbol{\alpha}}$ is given by $\hat{\boldsymbol{\alpha}} = \sum(u) \chi(u)^T \mathbf{W}(u) \tilde{\mathbf{Y}}$, where $\mathbf{W}(u) = \text{diag}\{K_h(\boldsymbol{\eta}^T \mathbf{Z}_1 - u), \dots, K_h(\boldsymbol{\eta}^T \mathbf{Z}_n - u)\}$, $K_h(\cdot) = K(\cdot/h)/h$, $\tilde{\mathbf{Y}} = (\tilde{Y}_1, \dots, \tilde{Y}_n)^T$, $\sum(u) = \{\chi(u)^T \mathbf{W}(u) \chi(u)\}^{-1}$ and $\chi(u)$ are the $n \times 2(p+1)$ matrix

$$\chi(u) = \begin{bmatrix} 1 & \tilde{X}_{11} & \dots & \tilde{X}_{1p} & (\boldsymbol{\eta}^T \mathbf{Z}_1 - u) & (\boldsymbol{\eta}^T \mathbf{Z}_1 - u) & \dots & (\boldsymbol{\eta}^T \mathbf{Z}_1 - u) \tilde{X}_{1p} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \tilde{X}_{n1} & \dots & \tilde{X}_{np} & (\boldsymbol{\eta}^T \mathbf{Z}_n - u) & (\boldsymbol{\eta}^T \mathbf{Z}_n - u) & \dots & (\boldsymbol{\eta}^T \mathbf{Z}_n - u) \tilde{X}_{np} \end{bmatrix}$$

The local linear approach requires selection of the bandwidth h . We apply the generalised cross-validation (GCV) criterion [17, 18] to select the bandwidth, as was done in [15]. Briefly, for a given $\boldsymbol{\eta}$, $\hat{\beta}_r(\boldsymbol{\eta}^T \mathbf{Z})$ is linear in $\{\tilde{Y}_i\}_{i=1}^n$. Therefore, the fitted values, $\hat{\tilde{\mathbf{Y}}} = (\hat{\tilde{Y}}_1, \dots, \hat{\tilde{Y}}_n)^T$, where $\hat{\tilde{Y}}_i = \sum_{r=0}^p \hat{\beta}_r(\boldsymbol{\eta}^T \mathbf{Z}_i) \tilde{X}_{ir}$, are also linear in $\{\tilde{Y}_i\}_{i=1}^n$. This means that $\hat{\tilde{\mathbf{Y}}} = \mathbf{V}(h) \tilde{\mathbf{Y}}$, where $\mathbf{V}(h)$ is the $n \times n$ hat matrix. (The formula for $\mathbf{V}(h)$ can be found in [15] or [4].) The bandwidth h is selected to minimise the following GCV criterion, which is a function of the residual sum of squares $RSS = \|\tilde{\mathbf{Y}} - \hat{\tilde{\mathbf{Y}}}\|^2$,

$$\text{GCV}(h) = n^{-1} \text{RSS} [1 - n^{-1} \text{tr} \mathbf{V}(h)]^{-2}. \quad (11)$$

3.2 Updating the coefficients of the linear combination of covariates

Given the functions $\{\beta_r(\cdot)\}$, we can search for the coefficients, $\boldsymbol{\eta}$, by minimising the mean squared error criterion

$$M(\boldsymbol{\eta}) = \frac{1}{n} \sum_{i=1}^n \left\{ \tilde{Y}_i - \sum_{r=0}^p \beta_r(\boldsymbol{\eta}^T \mathbf{Z}_i) \tilde{X}_{ir} \right\}^2. \quad (12)$$

A one-step iterative estimation procedure, analogous to the one-step Newton–Raphson estimation, can be used to update $\boldsymbol{\eta}$, as was done in [4] (see also [19]). More precisely, the updated coefficient vector can be obtained as

$$\boldsymbol{\eta}_1 = \boldsymbol{\eta}_0 \ddot{M}(\boldsymbol{\eta}_0)^{-1} \dot{M}(\boldsymbol{\eta}_0), \quad (13)$$

where $\dot{M}(\cdot)$ and $\ddot{M}(\cdot)$ are the derivative and the Hessian matrix of $M(\cdot)$, respectively, and $\boldsymbol{\eta}_0$ is the initial vector. The above estimator for $\boldsymbol{\eta}$ is based on the approximation $0 = \dot{M}(\hat{\boldsymbol{\eta}}) \approx \dot{M}(\boldsymbol{\eta}_0) + \ddot{M}(\boldsymbol{\eta}_0)(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)$, where $\hat{\boldsymbol{\eta}}$ is the minimiser of equation (12). This approximation holds when the initial value $\boldsymbol{\eta}_0$ is ‘close’ to $\boldsymbol{\eta}$. Therefore, we carefully examine this aspect in our numerical studies in section 4, via a detailed analysis of the sensitivity to initial values.

We point out here that the existence of a function of the form $\sum_{r=0}^p \beta_r(\boldsymbol{\eta}^T \mathbf{Z}) \tilde{X}_r$ that minimises equation (12) was proven in [4] under mild conditions (see Theorem 1 of [4]). The uniqueness of $\boldsymbol{\eta}$ and $\beta_r(\cdot)$ also follow, if we choose $\|\boldsymbol{\eta}\| = 1$ and the first non-zero

component of $\boldsymbol{\eta}$ to be positive. These conditions were incorporated into the proposed MCAR estimation algorithm. However, we note that the uniqueness of $\boldsymbol{\eta}$ and $\beta_r(\cdot)$ separately are not critical for MCAR estimation, as long as $\beta_r(\boldsymbol{\eta}^T \mathbf{Z})$ is unique. This is because the targeted quantities of interest are the underlying regression parameters, $\boldsymbol{\gamma}$, not the varying functions $\{\beta_r(\cdot)\}$.

3.3 Outline of the iterative algorithm

We summarise the full algorithm to obtain the estimates of the underlying regression parameters under the MCAR model given by equation (10). Let $\boldsymbol{\eta}_0$ be the normalised initial value and $\{h_1, \dots, h_K\}$ be a sequence of bandwidth values.

The algorithm is summarised in four main steps below.

- (a) Specify the initial value $\boldsymbol{\eta}_0$, bandwidths $\{h_1, \dots, h_K\}$ and convergence criterion δ .
- (b) For each bandwidth $h_k (k = 1, \dots, K)$ iterate/repeat (b1)–(b3) below until the absolute difference in mean square errors, $|M(\boldsymbol{\eta}_1) - M(\boldsymbol{\eta}_0)|$, is less than δ , where $\boldsymbol{\eta}_1$ is the new/updated value.
 - (b1) Given initial value $\boldsymbol{\eta}_0$, estimate the varying coefficient functions $\{\beta_r(\cdot)\}_{r=0}^p$, using local linear regression, as described in section 3.1.
 - (b2) Given the varying functions $\{\beta_r(\cdot)\}_{r=0}^p$ (from b1) estimate/update the coefficients, $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_0 - \ddot{M}(\boldsymbol{\eta}_0)^{-1} \dot{M}(\boldsymbol{\eta}_0)$, as described in section 3.2.
 - (b3) Update the initial vector, $\boldsymbol{\eta}_0 \rightarrow \boldsymbol{\eta}_1$, and repeat b1–b2 until convergence: $|M(\boldsymbol{\eta}_1) - M(\boldsymbol{\eta}_0)| < \delta$. The updated $\boldsymbol{\eta}_1$ is normalised, *i.e.*, $\boldsymbol{\eta}_1 \rightarrow \boldsymbol{\eta}_1 / \|\boldsymbol{\eta}_1\|$, as described in section 3.2.
- (c) Denote the final estimated coefficient vectors, for bandwidths $\{h_1, \dots, h_K\}$, by $\hat{\boldsymbol{\eta}}_1, \dots, \hat{\boldsymbol{\eta}}_K$, respectively. To determine the choice of bandwidth, compute $\{\text{GCV}(h_k)\}_{k=1}^K$ given by expression (11) using $\hat{\boldsymbol{\eta}}_k$ and $\{\hat{\beta}(\hat{\boldsymbol{\eta}}_k^T \mathbf{Z}_i)\}_{r=0}^p$, for $i = 1, \dots, n$. The final selected estimates, denoted by $\{\hat{\beta}(\hat{\boldsymbol{\eta}}^T \mathbf{Z}_i)\}_{r=0}^p$, correspond to $\text{argmin}_{h_k} \{\text{GCV}(h_k)\}$.
- (d) Finally, compute the estimates $\hat{\boldsymbol{\gamma}}_r$ given by equation (10).

In the implementation, we standardise the covariates $\{\mathbf{Z}\}$ to have a sample mean 0 and covariance matrix \mathbf{I}_s and iterate until convergence or when the number of iterations exceeds 40. Also, to reduce the computation, we estimate the $\beta_r(\cdot)$ (step C) on 101 regular grid points on $[-2, 2]$ and used linear interpolation to obtain values of the functions on the interval. (The (estimation) approximation of $\beta_r(\cdot)$ based on the grid points relative to the estimation based on the full n data points are very similar.)

4. Simulation studies

The primary aim of the simulation studies here is to assess how well the MCAR estimator, $\hat{\boldsymbol{\gamma}}$, targets the vector of true underlying regression parameters, $\boldsymbol{\gamma}$. We examine the bias and mean square error (MSE) of the MCAR estimator relative to (a) a baseline and (b) a benchmark. The benchmark estimator is obtained by assuming that $\boldsymbol{\eta}$ is known (*i.e.*, the estimator given by equation (9)), which is consistent for $\boldsymbol{\gamma}$. The baseline estimator is obtained simply by using the initial vector, $\boldsymbol{\eta}_0$, without the iteration steps b1–b3 (*i.e.*, taking $\boldsymbol{\eta}$ to be $\boldsymbol{\eta}_0$). Thus, we expect the performance of the MCAR estimator to be between the baseline and the benchmark. Additionally, because the bias of the MCAR estimator will depend on the uncertainty due to the unknown coefficient vector, $\boldsymbol{\eta}$, a study of the performance of the MCAR estimator requires

attention to the sensitivity to the initial vector, η_0 . Therefore, we examine the performance of the MCAR estimator for different starting values in all simulation studies.

4.1 Simulation design

To study the numerical properties of the MCAR estimator, we used the following simulation design. The underlying regression model considered was $Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \gamma_3 X_3 + e$, where the parameters are $(\gamma_0, \gamma_1, \gamma_2, \gamma_3) = (3, -1, 1, 1.5)$. The error variable is $e \sim N(0, 1)$ and $\mathbf{X} = (X_1, X_2, X_3)^T \sim N_3(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$, with mean $\boldsymbol{\mu}_X = (3, 1, 2)^T$ and covariance matrix

$$\boldsymbol{\Sigma}_X = \begin{bmatrix} 0.81 & 0.50 & 0.30 \\ 0.50 & 1.0 & 0.90 \\ 0.30 & 0.90 & 2.03 \end{bmatrix}.$$

Thus, the predictor correlations are $\rho(X_1, X_2) = 0.5556$, $\rho(X_1, X_3) = 0.2357$ and $\rho(X_2, X_3) = 0.6364$. To simulate the distorted (observed) data, we consider the following distorting functions, $\psi(U) = 1 + U/4$, $\phi_1(U) = (U_2 + 1)/2$, $\phi_2(U) = (U^2 + 2)/3$ and $\psi_3(U) = 1 + U^3/50$, where $U \equiv \boldsymbol{\eta}^T \mathbf{Z}$. (The constants in the above distorting functions were chosen to satisfy the identifiability constraints (6), specifically $E\{\psi(U)\} = 1$ and $E\{\psi_r(U)\} = 1$.) Thus, the distorted (observed) response and predictors are $\tilde{Y} = \psi(U)Y$ and $\tilde{X}_r = \psi_r(U)X_r$, as given by equation (5). Under this simulation setting, we considered the following four main cases for the distribution and dimension of the covariates, $\mathbf{Z} = (Z_1, \dots, Z_s)^T$.

Case 1 Two independent covariates, Z_1 and Z_2 , uniformly distributed on $[-1, 1]$ (denoted $Z_i \sim U[-1, 1]$) were considered.

Case 2 For this case, we considered two dependent covariates, $\mathbf{Z} = (Z_1, Z_2)^T$, distributed as bivariate normal with mean $\boldsymbol{\mu}_Z = (0, 1)^T$, correlation $\rho(Z_1, Z_2) = 0.80$, $\text{var}(Z_1) = 1$ and $\text{var}(Z_2) = 0.5$.

Case 3 Case 1, repeated in three dimensions, *i.e.*, $\mathbf{Z} = (Z_1, Z_2, Z_3)^T$, with $Z_i \sim U[-1, 1]$.

Case 4 To examine the dependent covariates in three dimension, we considered $\mathbf{Z} \sim N_3(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z)$, where $\boldsymbol{\mu}_Z^T = (0.5, 0, 1)$ and

$$\boldsymbol{\Sigma}_Z = \begin{bmatrix} 0.3333 & 0.4041 & 0.1225 \\ 0.4041 & 1.0000 & 0.2828 \\ 0.1225 & 0.2828 & 0.5000 \end{bmatrix}.$$

For the two and three distorting covariate cases, the unknown coefficient vectors to be estimated are $\boldsymbol{\eta} = (0.3162, 0.9487)^T$ and $\boldsymbol{\eta} = (0.1162, 0.3487, 0.9300)^T$ ($\|\boldsymbol{\eta}\| = 1$), respectively. In each simulation study, we generated 200 Monte Carlo data sets. For each generated data set, we obtained the MCAR, baseline and benchmark estimates. We considered sample sizes of $n = 250, 350, 550$ and 750 . Also, to satisfy the assumption of bounded support for \mathbf{Z} [2], we used a normal distribution truncated at ± 2.5 standard deviation for cases 2 and 4. (The results are similar for truncation at ± 3 standard deviation.)

We take the sequence of bandwidths $\{h_1, \dots, h_K\}$ to be $\{0.180, 0.200, 0.240, 0.288, 0.346, 0.415, 0.498, 0.597, 0.7170.860, 1.032\}$, which spans between 0.18 and 1.03 standard deviation of the covariate data $\{\mathbf{Z}_i\}$. The set of bandwidth values was generated to cover a range of

the standard deviation of the data (U) from the sequence $0.2(1.2)^{k-1}$ for a sequence of integer $k = 1, \dots, K$. We initially explored the sequence of bandwidth $\{0.10, 0.12, \dots, 2.67\}$ (*i.e.*, $0.1(1.2)^{(k-1)k} = 1, \dots, 19$), which covers 0.1 to 2.67 times the standard deviation of the data. However, preliminary simulation study suggests that GCV did not select bandwidths greater than about 1 standard deviation of the data. Thus, we reported the simulation study results using the reduced bandwidth sequence $\{0.18, \dots, 1.032\}$. For the local linear estimators of the varying functions, we used the Epanechnikov kernel $K(t) = 0.75(1 - t^2)_+$. In searching for η , we stopped the main iteration (steps b1–b3) if the convergence criterion was reached or the number of iterations exceeded 40 (for each bandwidth). The convergence criterion is met if the absolute difference between successive values of mean square errors, $|M(\eta_1) - M(\eta_0)|$, is less than 0.001.

4.2 Experimental results: numerical properties

For a given initial coefficient vector η_0 , the angle between η_0 and the true coefficient vector η is $\theta_0 = (180/\pi) \cos^{-1}(\eta_0^T \eta)$. We call θ_0 the ‘initial’ angle corresponding to the initial vector η_0 . To study the sensitivity of the MCAR estimator, $\hat{\gamma}$, to the initial vector, η_0 , we implemented simulation studies for various different starting vectors with initial angles ranging from 15° to 120° . We included in this range the ‘neutral’ initial vector given by $\eta_0^T = (1/\sqrt{s}, \dots, 1/\sqrt{s})_{1 \times s}$, which is the vector $(c, \dots, c)^T$ normalised (for any positive constant c). For two covariates, the neutral initial vector $\eta_0 = (0.7071, 0.7071)^T$ is about $\theta_0 = 27^\circ$ from η . Similarly, for three covariates, $\eta_0 = (0.5774, 0.5774, 0.5774)^T$ is about $\theta_0 = 36^\circ$ from η . Table 1 displays twelve initial starting vectors (and angles) corresponding the two- and three-dimensional cases we examined in the simulation studies.

Estimates and mean square errors (MSEs) for the MCAR, the baseline and the benchmark estimator are displayed in figure 1 for simulation Case 1 ($\mathbf{Z} = (Z_1, Z_2)^T \sim U[-1, 1]^2$). Displayed are results corresponding to the six different starting vectors η_0 . Each value plotted is an average over the 200 simulation runs, each with a sample size of $n = 350$. The left column of the four plots (top-down) corresponds to the estimates of $\gamma_0 = 3$, $\gamma_1 = -1$, $\gamma_2 = 1$, and $\gamma_3 = 1.5$. As expected, the benchmark estimator (which uses the true η) is closest to the true underlying regression parameters ($\gamma_r, r = 0, \dots, 3$). Given a ‘good’ starting value, for instance the η_0 corresponding to $\mu_0 = 15$, the MCAR estimator performs well. In this case, the MCAR estimates, $\hat{\gamma}$, are indistinguishable from the benchmark. For all six starting values considered, the MCAR estimator improves substantially (*i.e.*, has lower bias) relative to the baseline estimator. However, as anticipated, the performance of MCAR does vary, depending

Table 1. Starting values. Given are the initial vectors, η_0 , used for the two and three distorting covariates cases. Also given are the corresponding angle between each initial vector and the true vector η . Note that the initial vectors corresponding to ‘70a’ and ‘70b’ have an initial degree of 70, but the orientations or directions are different.

Two covariates		Three covariates	
Initial vector η_0^T	θ_0	Initial vector η_0^T	θ_0
(0.0599, 0.9982)	15	(0.3743, 0.3119, 0.8733)	15
(0.7071, 0.7071)	27	(0.5774, 0.5774, 0.5774)	36
(0.9285, 0.3714)	50	(0.6880, 0.6192, 0.3784)	50
(0.9996, 0.0300)	70a	(0.9300, 0.3487, 0.1162)	70a
(−0.7809, 0.6247)	70b	(0.9389, −0.1539, 0.3078)	70b
(0.6627, −0.7489)	120	(0.4554, 0.4554, −0.7650)	120

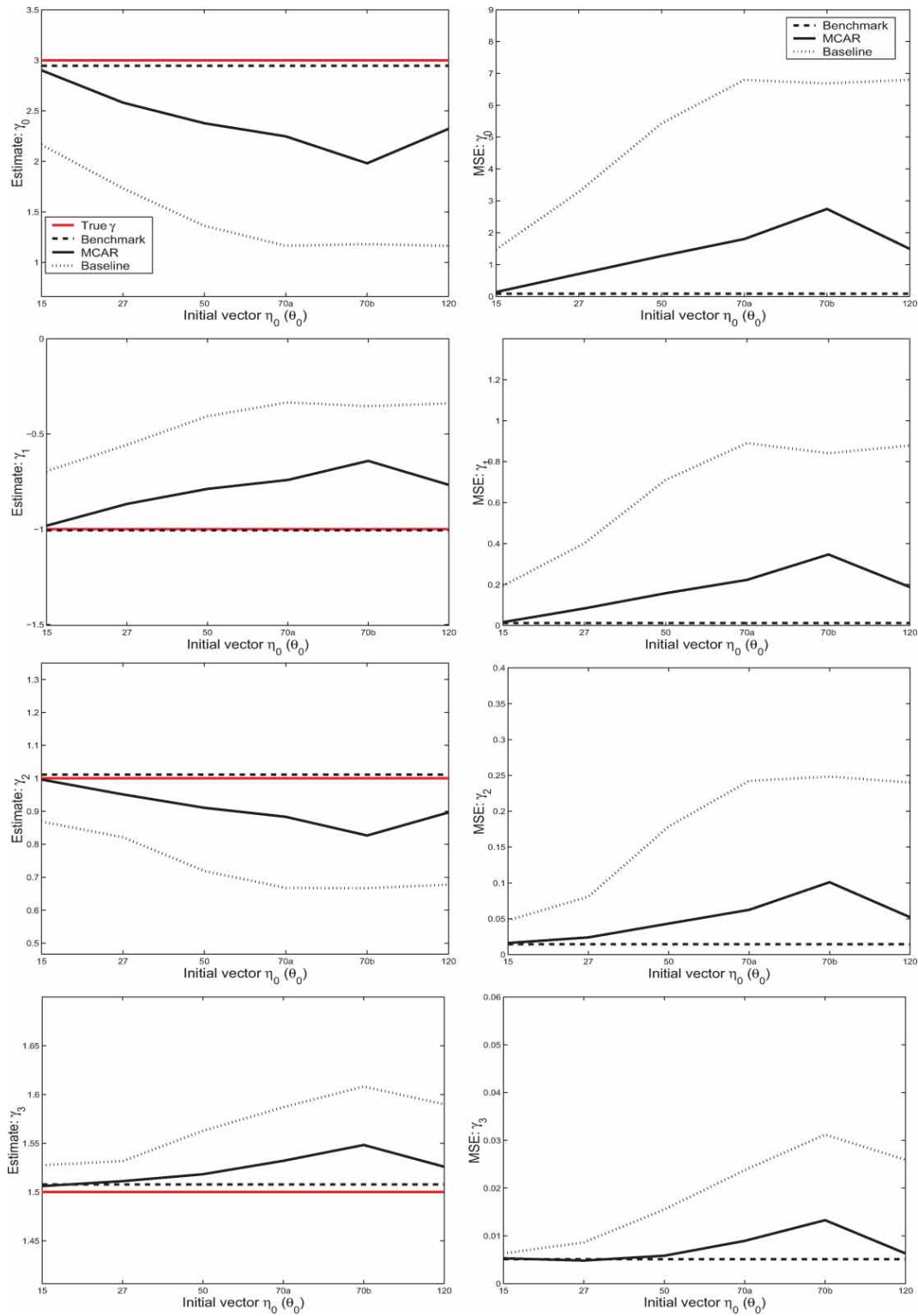


Figure 1. Case 1: Two independent (uniform) covariates. Given in the left column are the MCAR estimates (solid), $\hat{\gamma}$, the baseline (dotted) and benchmark (dashed) estimates. The true underlying parameters are indicated by the solid grey lines. The x-axis is the initial angle for each starting vector η_0 . The corresponding MSEs are given on the right column. Results are based on 200 Monte Carlo data sets ($n = 350$).

on the initial vector η_0 (corresponding to $\theta_0 = 15, 27, 50, 70a, 70b$ or 120). This is expected because the Newton–Raphson approximation (13), which inherently depends on the starting η_0 (as well as the sample size n , addressed in the next section). Also, we note that an increase in the angle (θ_0) between η_0 and the true coefficient vector η does not necessarily translate to a worst estimate of the underlying regression parameters. This is illustrated with the different results corresponding to $\theta_0 = 70a$ and $70b$. Although the angles for the two initial vectors are both 70° , the starting orientations (or directions) are different. Depending on the complexity of the surface $M(\eta)$, different starting orientations of η_0 can lead to different results.

The corresponding MSEs for simulation Case 1 are displayed in the right column of plots within figure 1. Since the benchmark estimator uses $\hat{\eta} \equiv \eta$, it does not depend on the starting value. Thus, the benchmark MSE is constant across θ_0 (as are the estimates, displayed in the left panel of figure 1). The MSE pattern suggests that the MCAR estimator clearly improves over the baseline and approaches the benchmark MSE, where the uncertainty regarding the unknown η has been eliminated. Also, for some initial vectors, the MSE for the MCAR estimator coincides with the benchmark MSE. The typical pattern of results, summarised for simulation Case 1 above, also holds for the two dependent (normally distributed) covariates (Case 2), for three independent uniform covariates (Case 3) as well as three dependent (normally distributed) distorting covariates (Case 4). The corresponding figures summarising the results for Cases

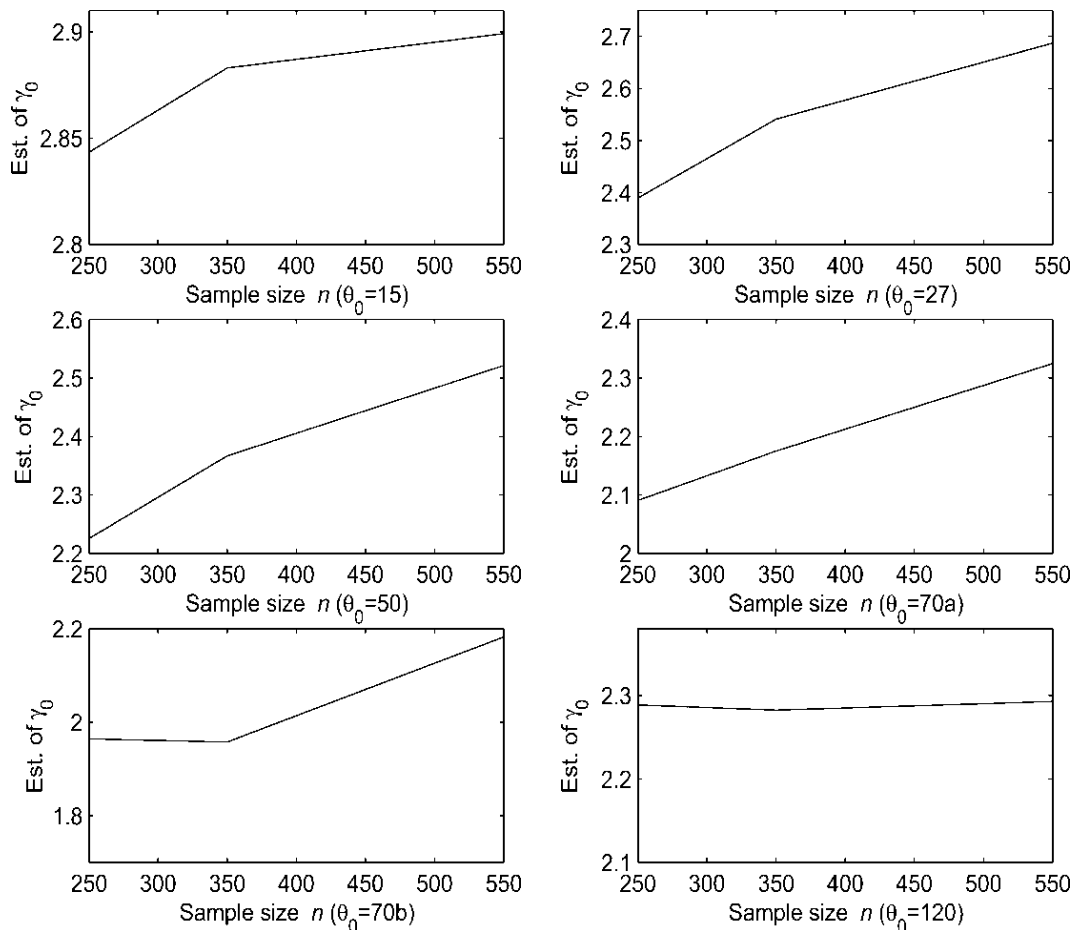


Figure 2. Sample size effect. Plotted are the MCAR estimates of $\gamma_0 = 3$ as a function of sample size n (x-axis). Generally, the MCAR estimates improve as n increases for each of the six starting vectors η_0 , corresponding to $\theta_0 = 15, 27, 60, 70a, 70b$ and 120 (simulation case 1).

2 through 4 are similar to figure 1 – therefore they are omitted here (and made available at <http://dnguyen.ucdavis.edu/.html/mcar.html>). However, we note the following two specific observations: (1) the bias and MSE are smaller for uniform covariates and (2) the bias and MSE are higher for the three-dimensional covariate \mathbf{Z} . This is not surprising because the search for $\boldsymbol{\eta}$, through minimisation of $M(\boldsymbol{\eta})$ using the Newton–Raphson procedure, can become increasingly challenging as the dimension increases.

We note that the $\text{Var}(\hat{\gamma}_r)$, estimated from the simulation runs, follows similar patterns as the MSEs given figure 1. With the uncertainty due to the unknown $\boldsymbol{\eta}$ removed, the benchmark estimator has the least variability and the (non-adaptive) baseline estimator has substantially higher variance than the MCAR estimator. For example, with three normally distributed covariates, the variances for estimating γ_3 corresponding to the baseline, MCAR and benchmark are 0.0222, 0.0125 and 0.0042 ($\theta_0 = 70a$), respectively. Thus, the variance of $\hat{\gamma}_3(\text{MCAR})$ is about 56% of the variance of $\hat{\gamma}_3(\text{baseline})$. Generally, the MCAR estimator has smaller variance than the baseline estimator.

Although the quality of the estimates, $\hat{\gamma}$, depends on the initial value $\boldsymbol{\eta}_0$, it should improve as n increases. We examined this property for the MCAR estimation algorithm. The results, summarised in figure 2, indicate that the MCAR estimates approach the true parameters as n increases. That is, the bias of the MCAR estimator decreases as the sample size n increases from 250 to 550. We observed this to be true for all initial values considered, except for the case corresponding to $\theta_0 = 120$ where the bias remained similar for sample sizes between 250 and 550. Displayed in figure 2 are the results only for γ_0 and for the six starting vectors (corresponding to $\mu_0 = 15, 27, 50, 70a, 70b$ and 120). Similar results were found for the other regression parameters $(\gamma_1, \gamma_2, \gamma_3)$ (results not shown). The median selected bandwidths corresponding to the six starting values were 0.2, 0.3456, 0.4147, 0.4147, 0.3456 and 0.4147, for the simulation Case 1 ($n = 350$). The median selected bandwidths for the other simulation cases were similar, ranging between 0.2 and 0.42.

5. Application to Pima Indian diabetes data

We illustrate the MCAR approach using a Pima Indian diabetes data set, which consists of 508 women at least 21 years old and of Pima Indian heritage. Briefly, patients with Diabetes Mellitus Type 2 may produce sufficient levels of insulin, but have abnormal insulin action (*e.g.*, insulin resistance) that prevents the body from normal utilisation of glucose. The problem of type 2 diabetes is also emerging in children and adolescents as well [20]. Typical chronic complications associated with diabetes are renal disease, loss of visual acuity, limb amputations and cardiovascular diseases, including hypertension. Obesity is a risk factor in both diabetes and hypertension.

To illustrate the methodology, we investigate the relationship between plasma glucose (glu) concentration and a hypertensive measure, diastolic blood pressure (dbp). In particular, we examine the following postulated underlying regression model between the response and predictor: $\text{glu} = \gamma_0 + \gamma_1 \text{dbp} + e$. Both the response and the predictor are potentially affected by various measurements of body configuration, including BMI ($Z_1 = \text{bmi}$) and triceps skin fold thickness ($Z_2 = \text{sft}$). Therefore, we directly adjust for these potential distorting covariates using the proposed MCAR method. The Pima Indian diabetes data set used here for illustration can be obtained at <http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>. We analysed the complete data available on $n = 508$ subjects 21 to 55 years old: response and predictors $\{\widetilde{\text{glu}}_i, \widetilde{\text{dbp}}_i\}_{i=1}^{508}$ and distorting covariates $\{(\text{bmi}_i, \text{sft}_i)\}_{i=1}^{508}$.

A regression of $\widetilde{\text{glu}}_i$ on $\widetilde{\text{dbp}}_i$, leads to the adaptive varying coefficient model, $\widetilde{\text{glu}}_i = \beta_0(U_i) + \beta_1(U_i)\widetilde{\text{dbp}}_i + \epsilon(U_i)$, where $U_i = \boldsymbol{\eta}_1 \text{bmi}_i + \boldsymbol{\eta}_2 \text{sft}_i$. One interpretation of U_i is

that it is a generalised index of obesity, which is relevant to both diabetes and hypertension, because obesity increases both insulin resistance and the risk of complications from high blood pressure. The estimated intercept and slope varying coefficient functions, $\hat{\beta}_0(\hat{U}_i)$ and $\hat{\beta}_1(\hat{U}_i)$, are displayed in figure 3 with selected index $\hat{\eta}^T = (0.5546, 0.8321)$ and bandwidth 1.14. The MCAR estimates of the underlying regression parameters adjust for the potential distorting effects of BMI and triceps skin fold thickness through the estimated varying coefficients. The MCAR estimates are $(\hat{\gamma}_0, \hat{\gamma}_1) = (92.955, 0.375)$. We estimate the standard errors of $\hat{\gamma}_0$ and $\hat{\gamma}_1$ based on 500 bootstrap samples. The corresponding standard error estimates are $\widehat{s.e.}(\hat{\gamma}_0) = 8.608$ and $\widehat{s.e.}(\hat{\gamma}_1) = 0.121$. The MCAR estimates and associated standard error estimates suggest that elevated levels of (diastolic) blood pressure is associated with increased plasma glucose concentration, even without the potential distorting effect of the generalised obesity index U .

The $\hat{\gamma}$ reported is based on the initial vector $\eta_0 = (0.7071, 0.7071)^T$. We also obtained various MCAR estimates based on different starting vectors to assess the stability of the MCAR estimates. In all starting values considered, the regression parameter estimates were similar; therefore, we take this as an indication that the given MCAR estimates are stable for the given data.

Next, based on the estimated varying coefficient functions $\hat{\beta}_0$ and $\hat{\beta}_1(\hat{U}_i)$, given in figure 3, we examine the form and type of distortion that the obesity index, U , has on the underlying plasma glucose and diastolic blood pressure. As mentioned in the Introduction section, without a priori knowledge of the specific form of the distortion, a simple approach to distortion adjustment is to divide by U . This approach assumes a special linear distortion of the form $\psi(U) = \psi(U) = U$ and that the distorting effect of the obesity index on plasma glucose and blood pressure are identical. If this assumption holds, then it follows that $\beta_1(U)$ is constant. Therefore, it is adequate to check to see if $\beta_1(U)$ is a flat horizontal line. The estimated $\hat{\beta}_1(U)$ in figure 3 suggests that this assumption may not hold and that the distortion effect of the obesity index on blood pressure is different from its effect on plasma glucose. Therefore, a simple adjustment via division by U is not justified. Additionally, the distortion effect of the obesity index on plasma glucose can be assessed directly from the estimated intercept function because $\beta_0(U) \propto \psi(U)$. The estimated intercept function $\hat{\beta}_0(U)$ in figure 3 suggests that the distortion on the response may be nonlinear in U . We also note that although the MCAR method adjusts for the distortion of U , whether the distortion is of a multiplicative, additive or no-distortion type, the specific type of distortion can also be assessed. That is, if the distortion effect of the generalised obesity index on plasma glucose and blood pressure is additive (i.e., $\widetilde{glu} = \psi(U) + glu$, $\widetilde{dbp} = \psi(U) + dbp$), then $\beta_1(U) = \gamma_1$ [2]. Again, the estimate $\hat{\beta}_1(U)$ is not constant; therefore, the results do not support an additive distortion model.

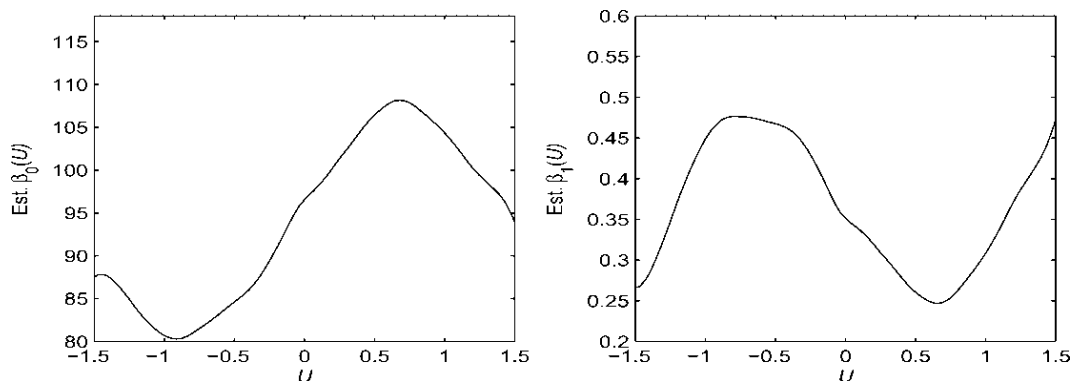


Figure 3. Estimates of varying coefficient functions, $\beta_0(u)$ and $\beta_1(u)$, from the Pima Indian data, corresponding to the intercept and the predictor variable dbp . The selected bandwidth is 1.14.

6. Discussion

The work presented here on MCAR extends the contaminated regression modelling capability by allowing for multiple distorting covariates, without restrictive assumptions on the exact form of the contaminating functions. Modelling the contaminations as general functions, although more complex, does have the added flexibility needed to adapt to varying levels of distortion complexities on the latent variables Y and X_r . However, despite the added modelling complexities, the interpretation of the latent variables remains the same. More precisely, they are defined as the parts of \tilde{Y} and \tilde{X}_r that are free from the effects of U . This interpretation applies to the simplest case of the univariate linear distortion model (1), where $U = Z$ and in the multicovariate distortion model (3), where $U = \beta^T \mathbf{Z}$. Additionally, if model (1), which assumes that the distorting functions are identity functions, is not correct then the adjusted variables will not be free from the effects of U . Instead, the division of \tilde{Y} and \tilde{X} by U in model (1) would lead to an artificial dependency between Y and X , since they would both be dependent on U by the division.

Additionally, by design, MCAR is robust to *distortion model misspecification*. Thus, an advantage of MCAR is that the estimation method targets the correct parameters under three distortion models: (1) additive, (2) multiplicative and (3) no distortion, as in the covariate adjusted regression approach of Sentürk and Müller [2]. Hence, the type of the distortion need not be known/specified for MCAR adjustment to be applicable. It is automatically adaptive to the above three types of distortion settings in that it will yield correct parameter estimates under all three settings. Also, the proposed framework allows for assessment of whether the distortion model can be reduced to an additive distortion, so that a simpler adjustment can be employed. To check whether the distortion is additive, it is sufficient to check whether the slope varying coefficient function is approximately a constant function. This is because, under an additive distortion on the response and predictor, the slope in the varying coefficient regression obtained from regressing the observed response on the observed predictor is constant.

Acknowledgements

We thank two referees for their suggestions that improved the paper. DVN is partially supported by grants from the NIEHS (ES013066 and ES011269).

References

- [1] Kaysen, G.A., Dubin, J.A., Müller, H.G., Mitch, W.E., Rosales, L.M., Levin, N.W. and the Hemo Study Group, 2003, Relationship among inflammation nutrition and physiologic mechanisms establishing albumin levels in hemodialysis patients. *Kidney International*, **61**, 2240–2249.
- [2] Şentürk, D. and Müller, H.G., 2005, Covariate-adjusted regression. *Biometrika*, **92**, 75–89.
- [3] Ichimura, H., 1993, Semiparametric least-squares, (SLS) and weighted SLS estimation of singleindex models. *Journal of Econometrics*, **58**, 71–120.
- [4] Fan, J., Yao, Q. and Cai, Z., 2003, Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society, Series B*, **66**, 57–80.
- [5] Cleveland, W.S., Grosse, E. and Shyu, W.M., 1991, Local regression models. In: J.M. Chambers and T.J. Hastie (Eds) *Statistical Models in S* (Pacific Grove: Wadsworth & Brooks), pp. 309–376.
- [6] Hastie, T. and Tibshirani, R., 1993, Varying coefficient models. *Journal of the Royal Statistical Society, Series B*, **55**, 757–96.
- [7] Ramsay, J.O. and Silverman, B.W., 1997, *The Analysis of Functional Data* (New York: Springer).
- [8] Nicholls, D.F. and Quinn, B.G., 1982, Random coefficient autoregressive models: an introduction. *Lecture Notes in Statistics*, **11**.
- [9] Chen, R. and Tsay, R.S., 1993, Functional-coefficient autoregressive models. *Journal of the American Statistical Association*, **88**, 298–308.
- [10] Hoover, D.R., Rice, J.A., Wu, C.O. and Yang, L.-P., 1998, Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809–822.

- [11] Chiang, C., Rice, J.A. and Wu, C.O., 2001, Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association*, **96**, 605–17.
- [12] Cai, Z., Fan, J. and Li, R., 2000, Efficient estimation and inferences for varying coefficient models. *Journal of the American Statistical Association*, **95**, 888–902.
- [13] Carroll, R.J., Fan, J., Gijbels, I. and Wand, M.P., 1997, Generalized partially linear single-index models. *Journal of the American Statistical Association*, **92**, 477–489.
- [14] Xia, Y. and Li, W.K., 1999, On single-index coefficient regression models. *Journal of the American Statistical Association*, **94**, 1275–1285.
- [15] Şentürk, D. and Nguyen, D.V., 2006, Estimation in covariate-adjusted regression. *Computational statistics and Data Analysis*, **50**, 3294–3310.
- [16] Fan, J. and Gijbels, I., 1996, *Local Polynomial Modelling and its Applications* (London: Chapman and Hall).
- [17] Wahba, G., 1977, A survey of some smoothing problems and the method of generalized crossvalidation for solving them. In: P.R. Krisnaiah (Ed) *Applications of Statistics* (Amsterdam: North Holland), pp. 507–523.
- [18] Craven, P. and Wahba, G., 1979, Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics*, **31**, 377–403.
- [19] Bickel, P.J., 1975, One-step Huber estimates in linear models. *Journal of the American Statistical Association*, **70**, 428–433.
- [20] American Diabetes Association, 2000, Type 2 diabetes in children and adolescents. *Diabetes Care*, **23**, 381–389.